programa final

Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais

INFORMAÇÃO PARA UMA SOCIEDADE MAIS JUSTA

III Conferência Nacional de Geografia e Cartografia

IV Conferência Nacional de Estatística

Reunião de Instituições Produtoras

Fórum de Usuários

Seminário "Desafios para Repensar o Trabalho"

Simpósio de Inovações

Jornada de Cursos

Mostra de Tecnologias de Informação

27 a 31 de maio de 1996
Rio de Janeiro, RJ BRASIL

IBGE 60 anos 1936-1996

Uma das maneiras de olhar o ofício de produzir informações sociais, econômicas e territoriais é como arte de descrever o mundo. Estatísticas e mapas transportam os fenômenos da realidade para escalas apropriadas à perspectiva de nossa visão humana e nos permitem pensar e agir à distância, construindo avenidas de mão dupla que juntam o mundo e suas imagens. Maior o poder de síntese dessas representações, combinando, com precisão, elementos dispersos e heterogêneos do cotidiano, maior o nosso conhecimento e a nossa capacidade de compreender e transformar a realidade.

Visto como arte, o ofício de produzir essas informações reflete a cultura de um País e de sua época, como essa cultura vê o mundo e o torna visível, redefinindo o que vê e o que há para se ver.

No cenário de contínua inovação tecnológica e mudança de culturas da sociedade contemporânea, as novas tecnologias de informação - reunindo computadores, telecomunicações e redes de informação - aceleram aquele movimento de mobilização do mundo real. Aumenta a velocidade da acumulação de informação e são ampliados seus requisitos de atualização, formato - mais flexível, personalizado e interativo - e, principalmente, de acessibilidade. A plataforma digital vem se consolidando como o meio mais simples, barato e poderoso para tratar a informação, tornando possíveis novos produtos e serviços e conquistando novos usuários.

Acreditamos ser o ambiente de conversa e controvérsia e de troca entre as diferentes disciplinas, nas mesas redondas e sessões temáticas das Conferências Nacionais de Geografia, Cartografia e Estatística e do Simpósio de Inovações, aquele que melhor enseja o aprimoramento do consenso sobre os fenômenos a serem mensurados para retratar a sociedade, a economia e o território nacional e sobre as prioridades e formatos das informações necessárias para o fortalecimento da cidadania, a definição de políticas públicas e a gestão político - administrativa do País, e para criar uma sociedade mais justa.

Simon Schwartzman
Coordenador Geral do ENCONTRO

Promoção

Fundação Instituto Brasileiro de Geografia e Estatística
IBGE
Fundação Instituto Brasileiro de Geografia e Estatistica
IBGE
Associação Brasileira de Estudos Populacionais
ABEP

Co-Promoção

Associação Brasileira de Estatística
ABE
Associação Brasileira de Estudos do Trabalho
ABET
Associação Brasileira de Pós-graduação em Saúde Coletiva
ABRASCO
Associação Nacional de Centros de Pós-graduação em Economia
ANPEC
Associação Nacional de Pós-graduação e Pesquisa em Ciências Sociais
ANPOCS
Associação Nacional de Pós-graduação e Pesquisa em Geografia
ANPEGE
Associação Nacional de Pós-graduação e Pesquisa em Planejamento Urbano e Regional
ANPUR
Sociedade Brasileira de Cartografia
SBC

Apoio

Federação das Indústrias do Estado do Rio de Janeiro
FIRJAN
Academia Brasileira de Letras
ABL
Conselho Nacional de Pesquisas
CNPq
Financiadora de Estudos e Projetos
FINEP
Revista Ciência Hoje

Institutos Regionais Associados

Companhia do Desenvolvimento do Planalto Central
CODEPLAN (DF)
Empresa Metropolitana de Planejamento da Grande São Paulo S/A
EMPLASA (SP)
Empresa Municipal de Informática e Planejamento S/A
IPLANRIO (RJ)
Fundação Centro de Informações e Dados do Rio de Janeiro
CIDE (RJ)
Fundação de Economia e Estatística
FEE (RS)
Fundação de Planejamento Metropolitano e Regional
METROPLAN (RS)
Fundação Instituto de Planejamento do Ceará
IPLANCE (CE)
Fundação João Pinheiro
FJP (MG)
Fundação Joaquim Nabuco
FUNDAJ (PE)
Fundação Sistema Estadual de Análise de Dados
SEADE (SP)
Instituto Ambiental do Paraná
IAP (PR)
Instituto de Geociências Aplicadas
IGA (MG)
Instituto de Pesquisas Econômicas, Administrativas e Contábeis
IPEAD (MG)
Instituto do Desenvolvimento Econômico Social do Pará
IDESP (PA)
Instituto Geográfico e Cartográfico
IGC (SP)
Instituto de Apoio à Pesquisa e ao Desenvolvimento "Jones dos Santos Neves"
IJSN (ES)
Instituto Paranaense de Desenvolvimento Econômico e Social
IPARDES (PR)
Processamento de Dados do Município de Belo Horizonte S/A
PRODABEL (MG)
Superintendência de Estudos Econômicos e Sociais da Bahia
SEI (BA)

Coordenação Geral

Simon Schwartzman

Comissões de Programa

Confege

César Ajara (IBGE)
Denizar Blitzkow (USP)
Jorge Marques (UFRJ)
Lia Osório Machado (UFRJ)
Mauro Pereira de Mello (IBGE)
Speridião Faissol (UERJ)
Trento Natali Filho (IBGE)

Confest

José A. M. de Carvalho (UFMG)
José Márcio Camargo (PUC)
Lenildo Fernandes Silva (IBGE)
Teresa Cristina N. Araújo (IBGE)
Vilmar Faria (CEBRAP)
Wilton Bussab (FGV)

Comissão Organizadora

Secretaria Executiva - Luisa Maria La Croix
Secretaria Geral - Luciana Kanham
Confege, Confest e Simpósio de Inovações
Anna Lucia Barreto de Freitas, Evangelina X.G. de Oliveira, Jaime Franklin Vidal Araújo, Lilibeth Cardozo R.Ferreira e Maria Letícia Duarte Warner
Jornada de Cursos - Carmen Feijó
Finanças - Marise Maria Ferreira
Comunicação Social - Micheline Christophe e Carlos Vieira
Programação Visual - Aldo Victorio Filho e Luiz Gonzaga C. dos Santos
Infra-Estrutura - Maria Helena Neves Pereira de Souza
Atendimento aos Participantes - Cristina Lins
Apoio
Andrea de Carvalho F. Rodrigues, Carlos Alberto dos Santos, Delfim Teixeira, Evilmerodac D. da Silva, Gilberto Scheid, Héctor O. Pravaz, Ivan P. Jordão Junior, José Augusto dos Santos, Julio da Silva, Katia V. Cavalcanti, Lecy Delfim, Maria Helena de M. Castro, Regina T. Fonseca, Rita de Cassia Ataualpa Silva e Taisa Sawczuk

Registramos ainda a colaboração de técnicos das diferentes áreas do IBGE, com seu trabalho, críticas e sugestões para a consolidação do projeto do ENCONTRO.

Planejamento, estimação e análise de dados em pesquisas por amostragem: desvendando a realidade brasileira com o "telescópio da estatística"

Pedro Luis do Nascimento Silva[1]

[1]IBGE - Departamento de Metodologia, atualmente afastado para estudos de doutoramento no Department of Social Statistics, University of Southampton, Southampton SO17 1BJ, Inglaterra.

Resumo

Este artigo apresenta uma revisão abrangente, embora superficial, dos desenvolvimentos ocorridos em termos do emprego de pesquisas por amostragem para obtenção, produção e análise de informações no IBGE, revendo também progressos recentes da teoria de amostragem. As três principais questões em termos de metodologia estatística para pesquisas por amostragem são: como selecionar amostras, como estimar as quantidades desconhecidas de interesse e como formular e ajustar modelos com os dados das amostras observadas. O artigo examina progressos feitos nessas questões e aponta desafios que o IBGE enfrentará nos anos vindouros para aproveitar plenamente o instrumento da amostragem em benefício do conhecimento preciso e detalhado da realidade brasileira.

Abstract

This paper presents a comprehensive but superficial review of developments in the use of sample surveys to produce and analyse information at IBGE, including a discussion of recent developments in sampling theory. The three key questions in terms of statistical methodology for sampling surveys are: how to select samples, how to estimate the target unknown quantities and how to pose and fit models given sample survey data. The paper reviews progress made on these questions and indicates challenges for IBGE in the coming years to fully benefit from the powerful instrument of sampling for a precise and detailed knowledge of the brazilian reality.

Agradecimentos

Minha maior gratidão a Denise B.N. Silva, Zélia M. Bianchini, Antonio J.R. Dias, Sonia Albieri, Fernando A.S. Moura, Mauricio T.L. Vasconcellos e Fernanda M. Santis, sem cujo apoio esse artigo não teria sido possível.

1. Introdução

As pesquisas por amostragem desempenham na Estatística papel semelhante ao do telescópio na Astronomia. Sua "invenção" em meados deste século e popularização na segunda metade do mesmo permitiram explorar com rigor científico vastas áreas do universo econômico e social, no Brasil e no resto do mundo, de forma antes não imaginada. Dados obtidos através de pesquisas por amostragem estão hoje na base de conhecimento de todas as ciências sociais.

Isso justifica estudar como o instrumento (pesquisas por amostragem) tem evoluído ao longo do tempo, e em particular como tem sido empregado no IBGE. Este artigo apresenta uma revisão abrangente, embora superficial, dos desenvolvimentos ocorridos no IBGE com relação à utilização de amostragem para obtenção e produção de informações e sua posterior análise. Ao mesmo tempo, procura rever progressos recentes da teoria de amostragem e indicar suas conexões e relevância para o IBGE e os usuários de suas informações. A amplitude da área a ser coberta é grande, mas a ambição do artigo é modesta. Pretende-se aqui destacar os principais aspectos e idéias, e não fazer uma descrição ou revisão detalhada dos mesmos.

Pesquisas por amostragem envolvem três principais questões em termos de metodologia estatística: como selecionar amostras, como estimar as quantidades desconhecidas de interesse (parâmetros) e como formular e ajustar modelos com os dados das amostras observadas. Essas questões orientam a divisão do artigo. O capítulo 2 contém uma discussão da evolução dos aspectos ligados ao planejamento de pesquisas por amostragem. Após breve revisão histórica da introdução da amostragem no IBGE, as modificações introduzidas em várias pesquisas são revistas e os progressos alcançados indicados.

Os aspectos ligados à estimação de totais, médias e proporções nas pesquisas por amostragem do IBGE são discutidos no capítulo 3, destacando-se o emprego de estimadores que incorporam informações auxiliares na estimação e alternativas para tratamento de não-resposta.

A análise de dados provenientes de pesquisas por amostragem é a questão tratada no capítulo 4. São apontadas dificuldades causadas porque os dados são gerados por desenhos amostrais complexos. Argumenta-se que a utilização automática de métodos da estatística clássica tais como análise de regressão e outros pode ser problemática. Algumas estratégias para lidar com o problema são apontadas bem como se argumenta que essa é uma área em que o IBGE e seus usuários tem muito que aprender e que fazer para superar as dificuldades.

O capítulo 5 destaca as conclusões do artigo e indica desafios que o IBGE enfrentará nos anos vindouros para aproveitar plenamente o poderoso instrumento da amostragem em benefício do conhecimento preciso e detalhado da realidade brasileira.

2. Planejamento de Pesquisas por Amostragem

A utilização de amostragem nas pesquisas do IBGE já tem uma história considerável. Iniciando-se com a coleta de dados por amostra durante o Censo Demográfico (CD) de 1960, o IBGE foi sendo progressivamente "conquistado" pelo uso de amostragem. A introdução da Pesquisa Nacional por Amostra de Domicílios (PNAD) em 1967 é um dos marcos dessa história (IBGE, 1981a), seguindo-se o Censo Demográfico de 1970 (IBGE, 1983b).

Nos censos de população, a amostragem foi introduzida primordialmente para reduzir custos de coleta de dados mais detalhados (questionário longo), cuja obtenção para a população como um todo era muito cara. Em 1970 empregou-se amostragem também na pesquisa de avaliação da qualidade, para preparar tabulações avançadas e para gerar arquivos de uso público (pequena amostra com registros individuais desidentificados para servir ao público interessado em desenvolver suas próprias análises). A realização do Estudo Nacional da Despesa Familiar (ENDEF) em 1974/75 marca mais uma etapa desse processo de consolidação no IBGE do emprego da amostragem como uma ferramenta indispensável ao processo de pesquisa e produção de informações (Vasconcellos, 1983).

Nessa primeira fase, o uso de amostragem pelo IBGE se restringiu a pesquisas em que o domicílio era a unidade de investigação. Nas pesquisas de instituições ou estabelecimentos, a norma era a realização de censos ou pesquisas exaustivas, tentando cobrir todas as unidades das respectivas populações de interesse. Uma exceção digna de nota foi a Pesquisa Industrial Mensal - Produção Física (PIM-PF) iniciada em 1971, a qual pesquisa repetidamente uma amostra intencional das maiores indústrias (IBGE, 1991a). Entretanto essa pesquisa não adota esquema probabilístico de seleção da amostra como as demais. Outra característica desse primeiro período foi a ênfase no emprego de assessoria externa para implementação da amostragem nas pesquisas do IBGE, levando os desenhos amostrais então adotados a se parecerem bastante com aqueles usados nas pesquisas congêneres de outros países.

No fim dos anos 70 e primeira metade dos anos 80 várias razões levaram o IBGE a ampliar bastante o uso de amostragem em suas pesquisas. Entre essas destacaram-se a maior demanda por informações ágeis sobre vários setores da economia e da sociedade, aumentos significativos dos tamanhos das populações alvo e a necessidade de pesquisar populações até então não cobertas.

Datam desse período a implantação do Sistema Nacional de Índices de Preços ao Consumidor[2] (SNIPC) em 1979 e da Pesquisa Mensal de Emprego[3] (PME) em 1980, o uso de amostragem em pesquisas da indústria, começando com a Pesquisa Especial da Indústria[4] de 1981,

---

[2]Veja IBGE (1994a).

[3]Veja IBGE (1983a).

[4]Veja IBGE (1982a).

prosseguindo com a Pesquisa Industrial Anual[5] (PIA) desde 1981, o Índice de Preços ao Produtor industrial[6] (IPP) de 1981 a 1985, e mais tarde com a reformulação da Pesquisa Industrial Mensal de Dados Gerais[7] (PIM-DG) em 1985. Vale mencionar ainda que no Censo Demográfico de 1980 a amostragem foi empregada de forma ainda mais ampla que em 1970, com a introdução de esquemas de controle de qualidade baseados em inspeção por amostragem durante as etapas de coleta e processamento.

Essa fase de consolidação da amostragem como instrumento básico para a realização de pesquisas foi caracterizada por maior independência e autonomia técnica. O planejamento amostral passou a ser realizado por pessoal do próprio IBGE, incorporando inovações técnicas e refletindo o desenvolvimento de uma cultura própria. Data ainda dessa fase o reconhecimento da necessidade de documentar e disseminar entre os usuários as metodologias das pesquisas, em particular detalhes dos procedimentos amostrais adotados, tendo sido publicados relatórios descrevendo as metodologias adotadas nas principais pesquisas do IBGE.

Foram várias as inovações introduzidas na área de planejamento amostral. Nas pesquisas por estabelecimentos vale mencionar o emprego de desenhos amostrais com probabilidades desiguais de seleção. A PIA-81 empregou o método de "Poisson sampling" (IBGE, 1982b) para selecionar respondentes produtores de uma lista de produtos pré-determinada a fim de investigar dados de produção física da indústria. Foi também empregada alocação ótima para distribuir a amostra entre os estratos na pesquisa de dados gerais. O IPP empregou amostragem sistemática com probabilidades desiguais para selecionar estabelecimentos nos quais seria observada a variação dos preços de produtos industriais.

Os desenhos amostrais das PIAs de 1983 e 1984 foram modificados para refletir mudanças nos objetivos da pesquisa. Foram utilizados um novo cadastro base, nova estratificação e introduzidas amostras complementares de novos estabelecimentos captados através de registros administrativos (IBGE, 1984; Cabral, 1985). Na implantação do SNIPC empregou-se amostragem para montar o cadastro e selecionar a amostra da Pesquisa de Locais de Compra, os quais seriam visitados para a pesquisa de preços (IBGE, 1994b), e também na Pesquisa de Especificação de Produtos cujos preços seriam monitorados (IBGE, 1994a).

Outra tendência desse período foi o uso de computadores para o planejamento, seleção e expansão de amostras. Programas usando o sistema SAS foram desenvolvidos para implementar as estratégias amostrais de várias pesquisas (Silva, 1989).

---

[5]Veja IBGE (1982b).

[6]Veja IBGE (1982c).

[7]Veja IBGE (1991a).

Essa fase de consolidação foi seguida por uma fase de aperfeiçoamento (após 1985), em que as estratégias amostrais de várias pesquisas foram revistas, buscando redução dos custos, maior eficiência, adequação a novos objetivos e incorporação de novos cadastros. Entre essas vale mencionar a redução da amostra da PNAD desde 1986 em cerca de 44% (Menezes et al., 1986), e da PME desde 1988 em 30% (Silva e Moura, 1988).

Também a PIA teve seu esquema amostral reformulado para o período 1986-1990. Foram introduzidas múltiplas unidades de amostragem (empresas e unidades locais) e amostragem de conglomerados (selecionando-se unidades locais e investigando-se todos os estabelecimentos nelas contidos).

Uma outra revisão amostral com grande impacto foi o emprego de duas frações amostrais no Censo Demográfico de 1991, a saber 10% em municípios com população projetada de mais de 15.000 habitantes e 20% nos demais (Silva e Bianchini, 1990). Essa modificação reduziu a fração global de amostragem empregada no Censo Demográfico de 25% para cerca de 12,5%, resultando em substantiva economia de recursos durante a coleta e processamento dos dados, sem que seja esperada redução apreciável da precisão das estimativas derivadas da amostra.

Nesse período, novas pesquisas empregando desenhos amostrais considerados modernos foram iniciadas, tais como a Pesquisa Anual do Comércio (PAC) em 1988 (IBGE, 1991b) e a Pesquisa de Orçamentos Familiares (POF) realizada em 1986/88 (IBGE, 1992). A PAC inovou pelo uso de dois cadastros distintos para seleção da amostra (Censo 1985 e RAIS[8]) para compensar problemas que cada um desses cadastros teria isoladamente. Foi também empregado esquema amostral com estratificação automática e otimização do tamanho amostral (Hidiroglou, 1986) e introduzido o uso de amostra reserva para compensar as elevadas perdas de coleta esperadas devido aos problemas cadastrais (Silva et al, 1995), além de se ter adotado a empresa como unidade de seleção e investigação.

Já a POF 86/88 inovou por ter uma amostra de domicílios desenhada sob medida, com um desenho em dois estágios apenas, incorporando estratificação de setores censitários pela renda além da localização geográfica, e com setores selecionados com reposição e probabilidades proporcionais ao tamanho no primeiro estágio e domicílios com equiprobabilidade no segundo estágio (IBGE, 1992). Pela primeira vez, a seleção dos domicílios a pesquisar em cada setor foi feita sem empregar amostragem sistemática, mas sim amostragem aleatória simples, com o sorteio efetuado por computador. O dimensionamento da amostra também incorporou previsão de perda por não-resposta, outra novidade em relação à tradição das demais pesquisas domiciliares.

Também data desse período a introdução da amostragem em pesquisas da área agropecuária, com a Pesquisa Agropecuária do Paraná em 1986/87. Essa pesquisa empregou amostragem de áreas, com sofisticado esquema de estratificação multivariada incorporando

---

[8]Relação Anual de Informações Sociais.

restrições de contiguidade (Muller, Silva e Villalobos, 1988). Essa pesquisa foi mais tarde extendida para 3 outros estados (São Paulo, Distrito Federal e Santa Catarina) e está planejada sua extensão para 4 outros (Mato Grosso do Sul, Goiás, Rio Grande do Sul e Minas Gerais).

O início da década de 90 marcou um período de retrocesso, provocado por aguda crise de financiamento das operações do IBGE. Tais problemas levaram ao cancelamento de pesquisas (PNAD em 1994) ou a drásticas reduções não desejadas nas amostras de pesquisas da área econômica (PAC e PIA desde 1992), etc.

Essa fase de crise foi superada e no momento estão em curso várias iniciativas que representam progresso renovado no uso de amostragem. Foi iniciado em 1995 um programa abrangente de reformulação das estatísticas econômicas, apoiado na ampliação do uso de amostragem. Pela primeira vez, a montagem de cadastros para a seleção de amostras das pesquisas econômicas periódicas será baseada numa pesquisa cadastral também por amostragem (IBGE, 1995). No passado esses cadastros eram obtidos dos censos econômicos qüinqüenais, que entretanto não foram realizados no início da década de 90 como previsto. Já a Pesquisa Cadastro 1995 tem como base a RAIS, procurando aproveitar os dados cadastrais desse registro administrativo para suprir a lacuna deixada pela falta dos censos. Numa segunda fase, as diversas pesquisas econômicas repetidas terão suas amostras revistas, passando a selecionar unidades da amostra-mestre da Pesquisa Cadastro.

Tal modificação implica em novos problemas e dificuldades, causadas pela falta das bases cadastrais ("benchmarks") antes geradas pelos censos. Mas ao mesmo tempo enseja oportunidades para avanços e modernização dos desenhos amostrais das várias pesquisas econômicas. Uma vantagem do uso da RAIS é que novos dados são coletados anualmente. Isso não implica, entretanto, que tenha diminuído a importância dos censos como fornecedores de "benchmarks" para as pesquisas periódicas.

Outra iniciativa recente foi a introdução da Pesquisa Mensal do Comércio (PMC) em 1995, a princípio apenas no Rio de Janeiro. Essa pesquisa inovou no emprego de métodos de estratificação ótima (Farias e Barbosa, 1995), por ser a primeira incursão do IBGE no setor de comércio com uma pesquisa de periodicidade mensal e pela parceria com setores da sociedade civil no planejamento e financiamento da pesquisa.

Foi ainda realizada em 1994 uma Pesquisa piloto de Economia Informal (PEI), limitada ao município do Rio de Janeiro, com desenho amostral similar ao da POF 86/88. Uma inovação nessa pesquisa foi que a operação de listagem dos setores selecionados compreendeu uma primeira etapa da pesquisa, chamada "screening", na qual foram identificadas as unidades pertencentes à população alvo das quais foi selecionada a amostra efetivamente pesquisada. Essa pesquisa deverá ser realizada em escala completa em 1996, cobrindo as áreas urbanas dos 200 maiores municípios brasileiros.

Também está em curso desde outubro de 1995 a coleta de dados da POF 95/96, cujo desenho amostral é bastante similar ao adotado na POF 86/88, com pequenos ajustes decorrentes do uso de nova base cadastral obtida do Censo Demográfico de 1991 (Bianchini e Vieira, 1996).

A Pesquisa Domiciliar sobre Padrões de Vida (PPV), cuja pesquisa piloto foi iniciada em março de 1996, também faz parte dessa série de iniciativas recentes em que a amostragem tem papel de destaque. Seu desenho amostral se assemelha àquele adotado na POF 86/88, embora com abrangência geográfica distinta (regiões Nordeste e Sudeste do país) e menor número de unidades selecionadas em cada estágio de amostragem (Albieri, Bianchini e Cardoso, 1995).

Apesar da maturidade do IBGE em relação ao emprego de amostragem em suas pesquisas e à questão do planejamento amostral, ainda há vários desafios a serem superados. Primeiramente há que terminar a "conquista" da área de estatísticas agropecuárias, onde ainda se faz pouco uso de amostragem, em comparação com outras áreas. Um forte candidato seria o Censo Agropecuário, onde o emprego de amostragem durante a coleta poderia garantir economias significativas de tempo e recursos, nos moldes do que ocorre no Censo Demográfico.

Há também muito que fazer na direção de incorporar melhor a estrutura longitudinal de pesquisas repetidas como a PME, a PIM-DG e a PIM-PF ao rever os respectivos desenhos amostrais. A introdução de procedimentos de rotação nas amostras das pesquisas da área econômica, por exemplo, poderia proporcionar redução do ônus dos respondentes de menor porte ao mesmo tempo que daria oportunidade para atualizações mais freqüentes das mesmas. Já no caso da PME, há razões para crer que reduções ainda maiores no tamanho da amostra seriam possíveis sem grande perda de precisão, bastando para isso que a estrutura longitudinal da pesquisa seja mais plenamente aproveitada na hora da estimação, como se vai discutir no próximo capítulo.

Um último desafio seria a montagem de sistemas mais estáveis de geração e atualização das bases cadastrais das pesquisas, menos dependentes de censos, como se esboça na área econômica com a introdução da pesquisa cadastro. Tal objetivo ainda está fora de alcance para toda a área de pesquisas domiciliares pela falta de uma base cadastral digitalizada (incluindo mapas) e/ou de listas de endereços domiciliares em meio magnético, tais como as existentes em outros países.

3. Estimação em Pesquisas por Amostragem

Este capítulo discute a questão da estimação em pesquisas por amostragem. A revisão do uso de amostragem no IBGE apresentada no capítulo 2 serve como pano de fundo para ajudar a compreender o desenvolvimento também observado no uso de métodos de estimação.

Na primeira fase do uso de amostragem nas pesquisas do IBGE os estimadores empregados para totais e médias baseavam-se na idéia de ponderar as observações amostrais pelo inverso das respectivas probabilidades de seleção e tinham em comum o objetivo de simplicidade de cálculo. Os estimadores para totais eram da forma

$$\hat{Y} = \sum_{i \in s} w_i y_i \qquad (1)$$

onde $\hat{Y}$ é o estimador de Horvitz-Thompson do total populacional de uma variável y qualquer, y_i é o valor dessa variável para a i-ésima unidade amostral, w_i é o peso dado à i-ésima unidade amostral, calculado como $w_i = \pi_i^{-1}$, sendo π_i a probabilidade de inclusão na amostra da i-ésima unidade, e a soma é sobre o conjunto s de todas as unidades da amostra.

Devido ao emprego de desenhos amostrais "autoponderados"[9] (no Censo Demográfico, PNAD e PME, por exemplo), a simplicidade dos estimadores era "assegurada" pelo desenho amostral. Com as probabilidades de seleção todas iguais à fração amostral $f=n/N$, onde n é o tamanho da amostra e N o tamanho da população, os pesos w_i são todos iguais ao inverso dessa fração ($w_i = N/n \; \forall i$) e a expressão do estimador de total em (1) se reduz a

$$\hat{Y} = (N/n)\sum_{i \in s} y_i = N\bar{y}$$

onde $\bar{y}$ é a média amostral da variável y. Nesse período a preocupação com a estimação e divulgação da precisão das estimativas era secundária.

Tal esquema se justificava pelo imperativo prático de efetuar os cálculos das estimativas sem suporte computacional sofisticado. A ocorrência de problemas práticos tais como não-resposta (possivelmente diferencial em relação a subgrupos da amostra) e desatualização da base cadastral à medida que o período de referência da pesquisa se afastava do último censo, dificultava o uso puro e simples de estimadores "autoponderados". Na PNAD, por exemplo, o tratamento dispensado às "novas construções" foi motivado em parte pela aderência ao uso de estimadores autoponderados (Menezes, Almeida e Bianchini, 1991). Na PME um dos problemas que levou à necessidade de redução da amostra em 1988 foi provocado pela tentativa de manter a autoponderação do desenho mediante uso de fração fixa de amostragem no segundo estágio (Silva e Moura, 1988).

A despeito desses esforços para manter os estimadores autoponderados, observou-se que tais estimadores apresentavam vício não desprezível. A saída foi adotar estimadores que "calibrassem" a pesquisa, de forma que as estimativas de totais da população em certas categorias coincidissem com totais obtidos de projeções demográficas independentes. Isso era feito mediante emprego de estimadores de razão (quando apenas o total global da população é usado) ou pós-estratificação (quando a calibração se dá para duas ou mais categorias da população), em comum com a prática adotada em vários outros países. Nesse caso, os pesos amostrais são dados pelo inverso da fração amostral observada em cada categoria, isto é, $w_i = N_h/n_h$ se a unidade i pertence ao pós-estrato h.

Estimadores de pós-estratificação são uma forma simples de incorporar informações auxiliares para melhorar as estimativas amostrais. Essa melhoria se dá pela redução de vício

---

[9]Assim chamados porque todas as unidades populacionais têm idêntica probabilidade de inclusão na amostra.

provocado por não-resposta ou desatualização cadastral, ou pela redução da variância das estimativas devido ao efeito da estratificação. Quando o número de categorias nas quais a amostra é pós-estratificada é pequeno, o emprego desses estimadores é relativamente simples. É desse tipo o esquema de estimação adotado nas amostras dos Censos Demográficos de 1960 e 1970. Na PNAD e na PME são empregados estimadores de razão simples, calibrando-se apenas para o total da população residente projetada. Estimadores de razão com totais de pessoal ocupado e valor da produção foram também adotados na expansão da PIA de 1981 (IBGE, 1981b).

Uma vantagem óbvia da calibração é a concordância das estimativas da pesquisa com os totais de população projetados, sem prejuízo da simplicidade do processo de estimação. Entretanto, ela ajuda a mascarar problemas tais como a inexistência de tratamento adequado para a não-resposta (na PNAD e na PME), e as dificuldades causadas pela tentativa de aderência à autoponderação. Outra dificuldade é a qualidade das projeções de população usadas para calibrar a amostra (Bianchini, 1990). O Censo Demográfico de 1991 mostrou que as projeções populacionais para o final da década de 80 e início da década de 90 estavam superestimadas, e esse efeito foi repassado para as estimativas de totais originadas da PNAD e PME.

A incorporação de informações populacionais auxiliares na etapa de estimação de pesquisas por amostragem é a ligação comum a todas essas aplicações. Nesse assunto o IBGE já adquiriu larga experiência e tem feito, em algumas áreas, uso efetivo dos desenvolvimentos recentes da teoria. Para corroborar essa afirmação é preciso examinar a evolução dessa teoria e de como ela tem sido aplicada no IBGE.

O aproveitamento de informações populacionais auxiliares para estimação em pesquisas por amostragem é uma das partes da teoria de amostragem que mais progrediu desde os anos 70. O livro que representava o "estado da arte" da amostragem até então (Cochran, 1977) contemplava o uso de informações auxiliares através de estimadores de razão ou de regressão simples (ambos incorporando apenas uma variável auxiliar) ou de pós-estratificação. Entretanto essas técnicas eram apresentadas como ferramentas separadas, sem uma ligação comum.

O IBGE adotou na expansão da amostra do Censo Demográfico de 1980 o Processo Iterativo de Estimação por Totais Marginais (PIETOM) (IBGE, 1983b) aplicado separadamente para cada uma de 4219 áreas de ponderação[10]. Esse método consistia em definir uma tabela (ou matriz) de pós-estratificação de dupla entrada, cujas linhas e colunas eram dadas por combinações de valores das variáveis auxiliares, as quais foram investigadas a 100% através do questionário básico. Eram portanto conhecidos os totais populacionais das celas, linhas e colunas dessa tabela. Os pesos amostrais para unidades em cada cela eram calculados por um processo iterativo de ajuste dos pesos simples iniciais, de forma tal que as estimativas amostrais eram sucessivamente calibradas nos totais das linhas e depois das colunas, até que fosse observada convergência dos pesos.

---

[10]Área de ponderação é a menor área para a qual se calculava estimativas, e coincidia na maior parte das vezes com um município, podendo ser subdivisão deste nos de maior população.

O uso desse método permitiu ampliar bastante o número de variáveis auxiliares consideradas para a calibração das estimativas amostrais: a tabela de pós-estratificação empregada em 1980 tinha 720 celas, em comparação com os 46 pós-estratos adotados no Censo de 1970. No entanto, tal método envolvia um algoritmo iterativo que podia não convergir, requerendo ajustes na pós-estratificação quando isso ocorria. Além disso, o método foi implementado em programas desenvolvidos sob medida, requeria cálculos complicados para estimação da precisão e não oferecia uma teoria para explicar a definição da pós-estratificação. Essas dificuldades levaram o IBGE a procurar um novo método para expandir a amostra do Censo Demográfico de 1991, apesar do sucesso da aplicação do PIETOM em 1980.

Nesse período houve grande progresso na teoria. O livro que corresponde ao estado da arte da amostragem no início dos anos 90 (Särndal, Swensson e Wretman, 1992) apresenta as técnicas de pós-estratificação, estimação de razão e de regressão como casos particulares do estimador de regressão generalizado, o qual fornece uma estrutura flexível e eficiente para incorporar informações auxiliares na etapa de estimação. Esse estimador pode ser escrito na forma (1), mas nesse caso os pesos são dados por

$$w_i = \pi_i^{-1} g_i \tag{2}$$

onde g_i é um fator de ajuste ou calibração que incorpora informações sobre um vetor $x_i = (x_{i1}, \ldots, x_{iQ})'$ de variáveis auxiliares.

O estimador de regressão generalizado é motivado por um modelo linear relacionando a variável de interesse y com o vetor de variáveis auxiliares x tal como

$$y_i = \beta_0 + x_i'\beta + \varepsilon_i \tag{3}$$

onde β_0 e β são parâmetros desconhecidos, e se assume que os resíduos ε_i tem média 0 e são não correlacionados, com variância conhecida a menos de um fator de escala σ_e.

No caso mais simples em que a amostra é aleatória simples sem reposição e os erros do modelo (3) são supostos homoscedásticos, o fator de ajuste g_i é dado por:

$$g_i = 1 + (\overline{X} - \overline{x})' \hat{S}_x^{-1} (x_i - \overline{x}) \tag{4}$$

onde $\overline{X}$ é o vetor de médias conhecido das variáveis auxiliares na população, e $\overline{x}$ e $\hat{S}_x$ são respectivamente o vetor de médias e a matriz de covariâncias das variáveis auxiliares na amostra.

Com os fatores de ajuste dados por (4) é fácil mostrar que o estimador de regressão do total Y pode ser escrito como

$$\hat{Y}_R = N\left[\overline{y} + (\overline{X} - \overline{x})'b\right] \tag{5}$$

onde $b = \hat{S}_x^{-1}\hat{S}_{xy}$ é estimador de mínimos quadrados ordinários do parâmetro β e $\hat{S}_{xy}$ é a covariância amostral entre x e y.

O estimador de regressão é útil por pelo menos três motivos. Primeiro porque oferece calibração nas variáveis auxiliares, isto é, se aplicado a qualquer das variáveis do vetor x, replicará exatamente seu total conhecido na população. Segundo porque oferece ganhos de eficiência em relação ao estimador de Horvitz-Thompson sempre que a variância dos resíduos ε_i for menor que a dos y_i (Särndal, Swensson e Wretman, 1992, cap. 6). Terceiro porque tem grande flexibilidade, já que o vetor x de variáveis auxiliares pode incluir qualquer número de variáveis de tipo contínuo ou binário, ou ambos, e também devido ser facilmente generalizado para o caso de desenhos amostrais complexos.

Estimadores de regressão podem também ser justificados sob várias abordagens alternativas da teoria de amostragem. O livro de Särndal, Swensson e Wretman (1992) enfatiza uma abordagem "model assisted", em que o modelo de regressão (3) é usado para motivar o estimador, mas em que as propriedades do mesmo são avaliadas com respeito à distribuição gerada por repetidas replicações do processo de seleção da amostra.

Já Royall (1970) e seus seguidores apresentam esse estimador como o estimador ótimo sob uma família de modelos lineares, ignorando o desenho amostral na inferência e avaliando as propriedades do estimador com respeito a diferentes realizações do modelo postulado.

Mais recentemente, Deville e Särndal (1992) identificaram o estimador de regressão como um dos membros de uma família de estimadores de calibração da forma (1), em que os pesos w_i são da forma (2) com fatores de ajuste g_i obtidos de forma a minimizar a distância

$$D = E_P\left[\sum_{i\in s}\pi_i^{-1}(g_i-1)^2 \,/\, v_i\right] \tag{6}$$

sujeito à restrição de que $\sum_{i\in s}\pi_i^{-1}g_i x_i = \sum_{i\in U}x_i$, onde E_P representa uma esperança com relação à distribuição gerada pelo desenho amostral, e v_i são números positivos conhecidos, por exemplo proporcionais às variâncias dos erros ε_i no modelo linear (3). Empregando-se outras funções de distância se gera uma ampla família de estimadores que inclui os "raking ratio estimators" do método PIETOM, estimadores de regressão, de razão, de pós-estratificação e outros.

Estimadores de mínimos quadrados generalizados em dois estágios (MQG2) dessa família foram aplicados para a expansão da amostra do Censo Demográfico de 1991 (Silva, Bianchini e Albieri, 1993; Albieri e Dias, 1994). Eles substituíram o PIETOM com algumas vantagens. Esses

estimadores não dependem de processo iterativo de cálculo, são suportados por uma teoria sólida, simplificaram a estimação da precisão e forneceram recursos para automação do processo de seleção das variáveis auxiliares nas quais calibração pode ser imposta. Foi possível contar com programas mais ou menos gerais para implementação do método preparados e cedidos ao IBGE pelo Statistics Canada, cuja adaptação para uso no censo brasileiro foi feita com custo muito menor que o do desenvolvimento sob medida anterior.

Um dos usos habituais da amostra do censo demográfico é para estimação de medidas de concentração da renda. Essas medidas podem geralmente ser derivadas da função de distribuição da renda na população. Estimadores de pós-estratificação similares aos empregados na expansão da amostra do censo 1991 foram considerados em Silva e Skinner (1995) para estimar a função de distribuição populacional. Tais estimadores oferecem ganhos de precisão razoáveis em comparação com estimadores mais simples (Horvitz-Thompson, razão e diferença), sem necessidade de recorrer a estimadores não-lineares mais complexos disponíveis na literatura.

A ênfase dada aqui à discussão do método de estimação adotado no Censo Demográfico de 1991 se justifica porque dela se derivam vários outros aspectos de interesse. Primeiramente vale notar que tem aumentado no IBGE o emprego de sistemas genéricos em substituição a sistemas desenvolvidos sob medida, em resposta às necessidades de redução de custos e aumento da confiabilidade e eficiência das etapas de processamento dos dados (Silva e Bianchini, 1995). Na área de estimação em amostragem, em particular, há hoje em dia várias opções de sistemas genéricos: SUDAAN (Shah et al., 1992), GES (Estevao, Hidiroglou e Särndal, 1995), BASCULA (Göttgens et al., 1991), PC-CARP (Fuller, 1986), CALMAR (Sautory, 1993), CLAN (Andersson e Nordberg, 1994), WESVARPC (Westat, 1995). Todos esses sistemas são capazes de calcular estimativas de totais e médias, e respectivas medidas de precisão (exceto CALMAR), para uma ampla gama de desenhos amostrais e tipos de estimadores.

Em particular, o sistema GES desenvolvido no Statistics Canada implementa a metodologia de estimadores de regressão generalizados tal como descrita no livro de Särndal, Swensson e Wretman (1992). Embora a disponibilidade de tais sistemas não fosse a mesma no passado, o que ajuda a entender a prática do desenvolvimento local sob medida, eles não podem mais ser ignorados. Por justiça, vale mencionar que a estimação do ENDEF foi feita empregando um sistema genérico denominado ARIEL, à época considerado o estado da arte. No futuro, o IBGE deverá basear a estimação de suas pesquisas no emprego de sistemas genéricos desse tipo.

O método adotado no Censo 91 incorpora tratamento de não-resposta por reponderação automaticamente. Apesar da existência de outras opções para o tratamento de não-resposta, é consensual no mundo das estatísticas oficiais que não-resposta a nível de unidades perdidas por completo ("unit nonresponse") em pesquisas domiciliares deve ser tratada mediante reponderação. Portanto é um bônus contar com um método de estimação que tenha essa facilidade embutida.

Para não-resposta de itens o tratamento mais comumente aplicado é a imputação, cuja discussão está fora do escopo desse artigo. As experiências e progressos do IBGE nesse assunto foram revistas em Silva e Bianchini (1995).

Nas pesquisas domiciliares tradicionais do IBGE (PNAD e PME) não se emprega um tratamento adequado para não-resposta de unidades perdidas por completo. Já na POF 86/88 foram empregadas correções baseadas no inverso da taxa de não-resposta de cada setor da amostra, ao mesmo tempo que estimadores de razão baseados na projeção da população residente (IBGE, 1992). Nas pesquisas da área econômica a prática adotada tem sido a de reponderar as unidades respondentes de cada estrato para compensar a perda por não-resposta, como é o caso da PIA (IBGE, 1983c) e da PAC (IBGE, 1991b).

Apesar de se tratar de assunto no qual o IBGE já acumula alguma experiência, há grande desigualdade no reconhecimento do problema pelas suas diversas áreas. Há que investir mais no emprego de técnicas adequadas para monitoramento e tratamento da não-resposta, bem como na prática de reportar junto dos resultados os níveis de não-resposta observados nas várias pesquisas, para permitir avaliação pelos usuários dos dados. Nessa questão a POF 86/88 também deu exemplo, reportando dados detalhados de não-resposta total e por tipo. Mas a prática ainda não é adotada amplamente.

A metodologia MQG2 adotada no censo 1991 permite incorporar grande número de variáveis auxiliares, mas não oferece uma teoria para a escolha ótima das mesmas. Esse é um dos aspectos do emprego de estimadores de regressão que tem merecido atenção da comunidade de pesquisa recentemente. Em particular, Silva e Skinner (1996) apresentam um método para seleção de variáveis auxiliares quando se utiliza estimadores de regressão cuja eficiência para estimar a média de uma variável resposta especificada foi maior que a de vários competidores. Silva e Skinner (1996) apontam ainda para perda de precisão do estimador de regressão quando o número de variáveis auxiliares cresce demasiadamente, alertando para a necessidade de estabelecer um compromisso entre a calibração no maior número possível de variáveis auxiliares sem impor grande perda de eficiência no estimador.

O método MQG2 fornece estimativas de precisão de maneira mais simples que o PIETOM. Essa foi uma das áreas em que o IBGE progrediu bastante. Desde meados dos anos 80 passou a ser rotina o cálculo e disseminação de medidas da precisão das estimativas obtidas de pesquisas por amostragem, tanto as domiciliares como as da área econômica. A forma adotada para disseminar as estimativas de precisão tem variado no tempo e ainda representa um problema interessante.

Algumas vezes, como na PIA 1981, optou-se por duplicar o plano tabular publicando para cada tabela com estimativas de totais uma outra com estimativas dos respectivos coeficientes de variação (CVs). Essa opção dá estimativas de precisão em grande detalhe, mas sua implementação é cara. Mais recentemente o padrão tem sido publicar medidas agregadas da qualidade das estimativas, tais como tabelas resumindo as distribuições de freqüência dos CVs das estimativas de

total, como é o caso da PAC. Essa forma tem baixo custo, mas não permite conhecer o valor do CV de uma dada estimativa de total.

Um meio termo foi encontrado para disseminar a precisão das estimativas da amostra do Censo Demográfico de 1980, no qual foram empregadas funções de variância generalizadas ("generalized variance functions", Wolter, 1985) para permitir que usuários dos dados calculassem eles mesmos a precisão aproximada de qualquer estimativa de total desejada. Essas funções foram obtidas mediante um ajuste de modelos de regressão relacionando a precisão (CV) com o valor da estimativa (IBGE, 1983b). Essa abordagem foi adotada também na PNAD desde 1983. Já a PME não divulga estimativas de precisão, embora sejam calculados para uso interno os CVs estimados de algumas das taxas e totais divulgados com base nessa pesquisa. Uma solução "definitiva" para o problema ainda não está disponível, e esse é um tema que deve merecer atenção nos esforços de pesquisa metodológica do IBGE.

Um outro tema conectado com o método de estimação adotado no Censo 91 é a questão da estimação para pequenos domínios. Até o Censo de 1980, qualquer estimativa para domínios (áreas) menores que uma área de ponderação dependia de um processo de estimação caro e demorado, o que frustrava muitos usuários ou mesmo desencorajava um uso mais pleno dos dados do censo. O mesmo não pode ser dito do método adotado em 1991. Como a calibração é feita em dois níveis, um deles considerando totais populacionais por setor, é possível gerar estimativas para agregados de setores dentro de uma mesma área de ponderação utilizando os pesos calculados pela metodologia MQG2 adotando-se técnicas usuais de estimação em subpopulações. Se esses domínios tiverem tamanho mínimo da ordem de 5.000 habitantes, é provável que as respectivas estimativas tenham precisão aceitável.

Portanto não deverá mais ser necessário recorrer a métodos "ad hoc" para gerar estimativas para "agregados de setores" a partir da amostra do Censo 1991. Isso não elimina a possibilidade de que essas estimativas possam ser melhoradas usando métodos mais sofisticados, capazes de incorporar outras informações auxiliares. A questão de como estimar para pequenos domínios em pesquisas por amostragem foi outra em que a teoria da amostragem avançou bastante desde os anos 70. Na preparação de tabulações avançadas do Censo 1980 foram empregados estimadores sintéticos (Garcia, 1986) para gerar as estimativas a nível de microrregião (IBGE, 1983b).

Mais recentemente, Moura (1994) estudou o emprego de modelos hierárquicos para obtenção de estimadores para pequenos domínios. Esses estimadores se revelaram superiores quando comparados com estimadores convencionais encontrados na literatura mediante estudo de simulação usando dados do Censo Demográfico Experimental de Limeira de 1988, indicando que há ganhos quando se incorpora a estrutura hierárquica dos dados na modelagem e estimação. Apesar desses esforços, ainda há pouca aplicação prática no IBGE das modernas técnicas disponíveis para obtenção de estimativas para pequenos domínios. Essa é uma área em que investimentos relativamente modestos permitiriam aumentar substancialmente o valor dos dados das pesquisas através da divulgação de estimativas para áreas geográficas mais detalhadas, por exemplo.

Associada com a digitalização da base geográfica e o emprego de sistemas de informações geográficas (SIG), tal iniciativa permitiria considerar a dimensão geográfica na análise dos resultados das pesquisas de forma bem mais detalhada, o que enriqueceria bastante as possibilidades analíticas.

Métodos baseados no estimador de regressão podem também ser usados para combinar resultados de pesquisas distintas, a fim de aumentar a precisão das estimativas. Por exemplo, as pesquisas mensais poderiam ter suas estimativas calibradas anualmente à luz dos resultados das pesquisas anuais, cujas amostras (em geral bem maiores) fornecem estimativas mais precisas. Ou também pode ser empregada amostragem em duas etapas, como deverá ocorrer nas novas pesquisas da área econômica, que deverão ser subamostras da pesquisa cadastro (IBGE, 1995).

Um último aspecto da metodologia de estimação do Censo que se quer comentar aqui diz respeito aos pesos que ela gera para produção das estimativas. Essa metodologia, como também a do Censo 80 e de outras pesquisas como a POF 86/88, implica em que os pesos atribuídos aos registros individuais são "complexos", e não derivados da mera inversão das respectivas probabilidades de seleção. Isso tem implicações para os usuários de arquivos de microdados divulgados pelo IBGE, como se verá com mais detalhes no próximo capítulo.

A utilização de métodos mais elaborados de estimação nas pesquisas da área econômica é ainda incipiente. Nessa área há muito para fazer em termos de aproveitar melhor as informações auxiliares disponíveis e também em relação ao emprego de métodos capazes de compensar as deficiências da base cadastral existente. Isso sem falar que várias questões tratadas no âmbito das pesquisas domiciliares também se aplicam para essas pesquisas.

Encerrando esse capítulo, discute-se a questão da estimação em pesquisas repetidas, na qual pouco foi feito no IBGE. A grande maioria das pesquisas repetidas que o IBGE realiza ignora a estrutura longitudinal dos dados na hora da estimação, tratando cada rodada da pesquisa como se fosse uma pesquisa separada ("cross-section"). Isso é muito ineficiente se comparado com estimadores que combinam informações amostrais da rodada atual com as séries históricas armazenadas, como mostrado em Silva (1992) com uma aplicação a dados da PIM-DG. Tem havido grande progresso na teoria, com inúmeros artigos publicados recentemente, conforme a revisão em Silva (1995).

Trata-se de técnicas com grande potencial em termos de oferecer estimadores de melhor precisão sem necessidade de ampliar tamanhos de amostra. Ao contrário, provavelmente seu emprego propiciaria margem para redução dos tamanhos de amostra atuais. Entre as candidatas mais óbvias para emprego dessas técnicas estão a PIM-DG, PMC e a PME. Esta última poderia se beneficiar também de um redesenho que aproveitasse idéias como as empregadas na Pesquisa de Emprego e Desemprego na Grande São Paulo (PED) do SEADE/DIEESE, em particular a adoção de uma amostra trimestral com utilização de trimestres móveis para divulgação mensal de estimativas das taxas de interesse.

4. Análise de Dados de Pesquisas por Amostragem

A análise (estatística) de dados e resultados de pesquisas por amostragem é uma das áreas em que o IBGE tem muito que fazer e avançar. Entre o início dos anos 70 e meados da década de 80 grande importância era dada à questão da análise dos resultados das pesquisas no IBGE, embora essa análise nem sempre fosse apoiada na metodologia estatística. Esse período foi seguido por outro em que as prioridades foram alteradas e o IBGE passou a se concentrar mais na produção das informações básicas. Portanto não há hoje em dia uma tradição estabelecida de analisar em detalhe os resultados das pesquisas.

A prática de analisar os resultados das pesquisas enriquece as mesmas, aumenta o valor dos dados que elas produzem, fomenta o desenvolvimento técnico e contribui para o planejamento de pesquisas semelhantes no futuro. Em certa medida, também facilita um diálogo maior entre os usuários e o IBGE, pois este último passa a ter interlocutores mais qualificados e um entendimento melhor das necessidades dos primeiros. Por essas razões argumenta-se que o IBGE deve investir no desenvolvimento da sua capacidade de analisar dados e resultados das pesquisas, processo no qual uma parceria com a comunidade acadêmica e de usuários é indispensável.

Um outro motivo que justifica a realização de análises é o fato de que o IBGE tem acesso aos dados individuais detalhados de cada pesquisa, enquanto que os usuários freqüentemente tem acesso apenas a dados agregados, por razões ligadas à preservação do sigilo das informações. Como se verá mais adiante, essa é uma limitação séria para o emprego de métodos modernos de análise estatística. Isso sem mencionar que nem todos os usuários estão capacitados a processar grandes volumes de dados como os habitualmente produzidos pelas pesquisas do IBGE.

Tendo como objetivo ampliar a capacidade e o papel analítico do IBGE, é importante discutir como isso pode ser feito e que recursos a Estatística tem hoje para oferecer em suporte a essa tarefa. Para dar uma idéia resumida da evolução nessa questão, é interessante comparar dois cenários.

Por volta de 1970, os resultados das pesquisas por amostragem (ou censos) eram essencialmente tabelas com totais das variáveis pesquisadas por categorias de interesse. Essas tabelas eram condensadas em pesados livros ou "volumes de divulgação", provendo a única fonte de acesso aos dados por usuários externos. Um analista típico teria que compilar informações provenientes de vários desses volumes de tabelas e efetuar (manualmente) cálculos (necessariamente simples) para atingir conclusões. Como não se publicavam estimativas da precisão, as análises ficavam empobrecidas, pois não se tinha idéia da significância de efeitos ou diferenças encontradas nos dados. Comentários publicados junto das tabelas eram muitas vezes mera descrição das mesmas em palavras, pouco acrescentando em termos de uma compreensão dos resultados. Essa situação caracterizava essencialmente uma "análise descritiva (manual)" dos dados.

Hoje em dia ainda se publicam tabelas com totais das variáveis pesquisadas, mas também se coloca à disposição dos usuários arquivos com microdados (desidentificados) e mais recentemente

até acesso eletrônico direto às principais coleções, por exemplo via INTERNET. Um analista típico dispõe de um computador com acesso aos dados em formato eletrônico (local ou remoto) e de sistemas com sofisticados recursos para manejo dos dados, produção de gráficos e análise estatística. Com esses sistemas é possível analisar diretamente os dados individuais e formular, ajustar, testar e validar modelos levando em conta a precisão das estimativas e o fato de que os dados foram obtidos por amostragem. Análises desse tipo contribuem de maneira efetiva para a evolução do conhecimento, merecendo a descrição de "pesquisa" no sentido acadêmico da palavra.

Que fatores diferenciam o segundo cenário do primeiro? Em primeiro lugar, a disponibilidade de computadores e meios de comunicação cada vez mais poderosos e de baixo custo, garantindo a possibilidade de acesso direto aos dados pelos analistas interessados. Em segundo lugar, a ampliação do acesso aos microdados, mediante preparação de produtos de pronta entrega e de facilidades para acesso eletrônico remoto às bases de dados. Ainda há muito que fazer nessa área, mas há que registrar o progresso alcançado. Hoje o IBGE divulga rotineiramente arquivos de microdados de suas principais pesquisas domiciliares.

Além desses fatores, há que reconhecer a evolução da teoria Estatística, em grande parte sustentada pela evolução do poder computacional. No primeiro cenário, um analista interessado em investigar o diferencial de salários recebidos segundo o sexo do indivíduo ficaria limitado a comparar médias de salários dos respectivos grupos classificados por mais uma ou duas variáveis de controle, tais como setor de atividade ou idade. Tais comparações eram meramente descritivas, pois sem medidas da precisão das estimativas era impossível estabelecer a significância das diferenças encontradas.

Hoje em dia, dispondo dos dados individuais de uma pesquisa como a PNAD, por exemplo, um analista interessado na mesma questão pode formular modelos de regressão múltipla para explicar a variação dos salários em função de outras variáveis, e então testar a significância dos efeitos principais e interações. E é aqui que a evolução da teoria estatística se revela crucial. Se em meados dos anos 70 era comum o uso de modelos de regressão e outros modelos estatísticos em várias áreas de ciências, só mais recentemente essas técnicas foram adaptadas para aplicação na análise de dados provenientes de pesquisas por amostragem.

Na modelagem estatística convencional se assume que as observações de uma amostra[11] são independentes e identicamente distribuídas (IID). Tal hipótese é inadequada para o caso das pesquisas por amostragem realizadas para produção de estatísticas oficiais e muitas outras. Por razões práticas tais como inexistência de cadastros (de pessoas, por exemplo) ou de eficiência, essas pesquisas geralmente empregam desenhos amostrais "complexos" envolvendo estratificação, conglomeração e probabilidades desiguais de seleção. Embora adequados para gerar amostras cujas

---

[11]Aqui o termo é empregado no sentido amplo de um conjunto de observações de um certo fenômeno cujas propriedades se quer investigar.

estimativas de totais são representativas da população de interesse, esses desenhos produzem amostras que não podem ser adequadamente modeladas como IID.

Só recentemente se teve acesso amplo a uma abordagem teórica coerente e amadurecida para lidar com essas situações: o livro de Skinner, Holt e Smith (1989) discute métodos para a análise de dados de pesquisas complexas. Uma das idéias ali enfatizadas é a de que é essencial considerar o desenho amostral ao fazer inferência estatística empregando dados provenientes de amostras complexas.

Para ilustrar esse ponto, reconsidere o analista interessado em testar se há diferença significativa entre as médias dos salários recebidos por pessoas do sexo masculino e feminino, usando dados de uma amostra complexa. Se esse analista tentasse aplicar um teste t de Student usando um pacote estatístico qualquer, baseando sua análise na amostra de dados individuais, sua inferência estaria incorreta. Sem exceção, os procedimentos automatizados nos pacotes estatísticos de uso geral assumem que a amostra é IID. Isso geralmente implica em estimativas de precisão (essenciais para se fazer testes de hipótese e inferência estatística em geral) que são viciadas, subestimando grosseiramente o erro padrão efetivamente alcançado com a amostra.

Skinner, Holt e Smith (1989) propõe o uso de fatores de correção, denominados de "misspecification effects" (MEFF), para corrigir as estimativas de precisão habituais em razão da especificação incorreta da distribuição da amostra. O IBGE publicou estimativas de fatores desse tipo para algumas variáveis com base na amostra do Censo Demográfico de 1980 (Silva e Moura, 1990). Essa abordagem daria conta de resolver o problema para questões inferenciais relativamente simples, tais como comparações de médias por subgrupos ("ANOVA") e outras similares. Mas para a efetiva modelagem dos dados de pesquisas por amostragem essa abordagem não é suficiente, e vários capítulos do livro são dedicados a estudar alternativas.

Há dois caminhos principais: análises agregadas ou análises desagregadas. Na análise desagregada, os modelos postulados envolvem a especificação de distribuições para as variáveis de interesse na população alvo e para a relação dessas com variáveis auxiliares consideradas no desenho amostral. Nesse caso, as inferências podem ser feitas ignorando o desenho amostral, de vez que as variáveis auxiliares do desenho já estão diretamente incorporadas no modelo. A análise agregada requer apenas a especificação de modelos para descrever as distribuições das variáveis de interesse na população alvo, mas nesse caso as inferências para os parâmetros desses modelos devem ser feitas levando-se em conta o desenho amostral empregado.

Para ilustrar e analisar melhor as abordagens alternativas, considere o problema de ajustar um modelo de regressão linear usando dados provenientes de uma pesquisa por amostragem. Sejam y a variável resposta e z o vetor das variáveis explicativas de interesse no exercício de modelagem, e seja x o vetor de variáveis auxiliares consideradas no desenho amostral (por exemplo, variáveis indicadoras para os estratos de seleção ou para indicar a que conglomerado cada unidade pertence). O modelo de regressão linear relacionando y com z pode ser especificado alternativamente por (7) ou (8) a seguir:

$$y_i = \beta_0 + z_i'\beta + x_i'\gamma + \varepsilon_i \tag{7}$$

$$y_i = \beta_0 + z_i'\beta + \varepsilon_i \tag{8}$$

onde β_0, β e γ são parâmetros desconhecidos e os erros ε_i são não correlacionados, com média 0 e variância conhecida a menos de um fator de escala σ_e.

O modelo "desagregado" (7) incorpora diretamente as variáveis do desenho amostral. Nathan e Holt (1980), considerando o caso em que há apenas uma variável explicativa z e uma variável de desenho x, afirmam que 'não há nada de novo na situação: estimadores "model-based" não viciados dos parâmetros desse modelo estão disponíveis e a função do desenho amostral baseado na variável x é melhorar as propriedades desses estimadores'. Esse seria o domínio da estatística clássica, pois a modelagem explícita da resposta como função das variáveis auxiliares torna o desenho amostral não informativo (Smith, 1989).

Apesar de parecer simples, essa abordagem envolve questões complexas. Raramente a estrutura populacional considerada no planejamento amostral pode ser "codificada" numa única variável auxiliar. É mais comum haver dezenas de variáveis para representar os diferentes estratos e conglomerados aos quais pertencem as unidades. Isso motivou em parte o desenvolvimento recente no uso de modelos hierárquicos para analisar dados de pesquisas por amostragem (Goldstein e Silver, 1989; Bryk e Raudenbush, 1992), pois esses modelos permitem considerar a dependência entre unidades da amostra, incluindo efeitos de conglomeração.

Por outro lado, a inclusão de todas as variáveis auxiliares no modelo apresentaria desafios não triviais para o analista, em particular para a estimação adequada de todos os seus parâmetros. Além disso, a interpretabilidade e a validação do modelo se tornariam bem mais complexas (Skinner, Holt e Smith, 1989, p. 9). Outra dificuldade é que essa abordagem requer conhecimento detalhado sobre as variáveis x empregadas no desenho amostral, algo que freqüentemente não está disponível para o analista (secundário) devido a restrições de confidencialidade e outras razões práticas.

Modelos "agregados" como (8), nos quais as variáveis auxiliares não figuram diretamente, podem também ser preferidos por razões de substância: as variáveis de desenho x são em geral escolhidas por razões práticas e não porque fazem parte da explicação científica da resposta y pelas variáveis z. Apesar de não figurarem explicitamente no modelo (8), as variáveis auxiliares não podem ser ignoradas na estimação dos parâmetros do mesmo. Isso foi demonstrado de forma convincente por Holt, Smith e Winter (1980). Supondo que os vetores (y_i, z_i, x_i) têm distribuição normal multivariada, esses autores mostraram que o estimador de máxima verossimilhança do parâmetro β no modelo (8) depende da variável x.

A hipótese de normalidade não é válida em geral. Mas o que se verifica sob normalidade é que não se pode ignorar o efeito de seleção (Smith, 1993) das unidades populacionais com base nas variáveis auxiliares x a menos que uma condição seja satisfeita: (y_i, z_i) é independente de x_i na população. Como essa condição é raramente satisfeita e de difícil verificação, é portanto essencial considerar as variáveis auxiliares e/ou o desenho amostral na realização de inferências sobre β.

Holt, Smith e Winter (1980) e Nathan e Holt (1980) recomendam uma abordagem "model-based", que ignora o desenho amostral e se baseia no emprego de estimadores de Máxima Verossimilhança (MV) no modelo normal multivariado com dados incompletos. Binder (1983) propôs uma abordagem denominada de Pseudo Máxima Verossimilhança (PMV) que se aplica a modelos da família exponencial, mas que é bastante simples no caso do modelo linear (8) se os erros são normalmente distribuídos. Suponha que um censo fosse realizado. Nesse caso, o estimador B de MV de β baseado em toda a população finita seria obtido resolvendo-se a equação de verossimilhança

$$\sum_{i \in U} z_i (y_i - z_i' B) = 0 \quad \Leftrightarrow \quad (Z_U' Z_U) B = Z_U' Y_U \quad \Rightarrow \quad B = (Z_U' Z_U)^{-1} Z_U' Y_U \tag{9}$$

onde Z_U e Y_U são a matriz e o vetor de dados populacionais sobre as variáveis z e y respectivamente, e se assume que a matriz $Z_U' Z_U$ é inversível.

Como apenas uma amostra da população é observada, a idéia é usar os pesos amostrais w_i para estimar o total dos escores na população, e então obter o estimador de PMV de β mediante a solução das equações de pseudo verossimilhança:

$$\sum_{i \in s} w_i z_i (y_i - z_i' b_{PV}) = 0 \quad \Leftrightarrow \quad (Z_s' W_s Z_s) b_{PV} = Z_s' W_s Y_s \quad \Rightarrow b_{PV} = (Z_s' W_s Z_s)^{-1} Z_s' W_s Y_s \tag{10}$$

onde Z_s e Y_s são a matriz e o vetor de dados amostrais sobre as variáveis z e y respectivamente, W_s é a matriz diagonal com os pesos amostrais w_i, e a matriz $Z_s' W_s Z_s$ é suposta inversível.

Quando os pesos amostrais são todos iguais, o estimador em (10) coincide com o estimador de Mínimos Quadrados Ordinários (MQO) dado por

$$b_{MQ} = (Z_s' Z_s)^{-1} Z_s' Y_s \tag{11}$$

Já quando os pesos são os do estimador de Horvitz-Thompson, se obtém o estimador de Mínimos Quadrados Ponderados dado por:

$$b_\pi = (Z_s' \Pi_s^{-1} Z_s)^{-1} Z_s' \Pi_s^{-1} Y_s \tag{12}$$

onde Π_s é a matriz diagonal com as probabilidades de inclusão π_i.

O estimador MQO é viciado quando (y,z) é correlacionado com x e a seleção da amostra depende dessa variável (Nathan e Smith, 1989). Já o estimador de PMV é consistente (sob o desenho) para o parâmetro B definido em (9), embora haja perda de eficiência em relação ao estimador de MQO, que é "ótimo" sob o modelo. Essa troca entre vício e eficiência representa o leque de escolhas do analista na análise de dados de pesquisas por amostragem. Ignorar o desenho amostral pode implicar em vício considerável no estimador e respectivas estimativas de precisão. Para reduzir o vício considera-se o desenho amostral, mas há geralmente um preço a pagar em termos de eficiência. O debate sobre o papel que esses pesos amostrais devem desempenhar na inferência sobre modelos como (8) ainda não foi definitivamente resolvido (Smith, 1988; Pfeffermann, 1993). Uma coisa é certa, entretanto: ignorar pura e simplesmente o fato de que os dados provém de uma amostra complexa não é recomendável e os riscos dessa estratégia são grandes em termos da possibilidade de inferências incorretas.

Uma vantagem do método de PMV é sua relativa simplicidade e o fato de que depende de menos hipóteses que os métodos "model-based". Isso permitiu a sua implantação em sistemas genéricos como SUDAAN e PC-CARP, o que facilitou bastante a modelagem de dados amostrais complexos.

Recentemente, Silva (1996) tem estudado o efeito de empregar nas equações de pseudo verossimilhança os pesos do estimador de regressão dados por (2), ao invés dos pesos do estimador de Horvitz-Thompson. Esse estudo tem interesse prático porque muitas pesquisas divulgam dados individuais com pesos desse tipo, adotados para obter calibração ou para compensar não-resposta. Os resultados até agora apontam que esses pesos têm propriedades muito semelhantes aos de Horvitz-Thompson, embora haja cuidados a serem tomados em termos da estimação da precisão.

Essa discussão mostra a necessidade do IBGE prover seus usuários com informações suficientes para que possam fazer uso adequado e correto dos dados. Isso implica incluir pelo menos os pesos amostrais básicos (os inversos das probabilidades de inclusão) e variáveis indicadoras de estratificação e conglomeração junto dos microdados, ainda que por razões de sigilo estas variáveis talvez tenham que ser "codificadas".

Pela mesma razão se nota que o IBGE ocupa uma posição privilegiada em relação à possibilidade de efetuar análises dos dados. Provavelmente é o único capaz de experimentar com o uso de modelos desagregados como (7) devido às dificuldades causadas para analistas secundários pelas restrições de proteção do sigilo das informações. Isso é especialmente verdadeiro para estimação de precisão, pois esta requer informações mais detalhadas sobre o desenho amostral que aquelas necessárias para obtenção de estimativas pontuais.

A abordagem de PMV proposta por Binder (1983) é atrativa também porque se extende facilmente para outros modelos da família exponencial, tais como regressão logística, modelos log-lineares para análise de tabelas de contingência e outros modelos lineares generalizados. Alguns desses estão já implementados em sistemas genéricos como SUDAAN.

Essa facilidade e a maior disseminação de microdados levarão a um uso mais sofisticado dos mesmos. por sua vez gerando maior demanda por novos microdados. Atender satisfatoriamente essa demanda é mais um desafio para o IBGE, que precisa ao mesmo tempo cumprir sua responsabilidade de preservar o sigilo de informações individuais identificadas.

Outra área em que o progresso da teoria vai desafiar o IBGE é a de análise das pesquisas longitudinais. Já existe alguma tradição em pesquisas como a PIM-PF e PIM-DG de analisar as séries históricas agregadas e publicar séries sazonalmente ajustadas (IBGE, 1991a). Mas essa é uma análise bastante crua, e nem mesmo praticada em todas as pesquisas repetidas do IBGE. A PME, por exemplo, ainda não faz ajuste sazonal e só recentemente se tem investigado a obtenção de séries sazonalmente ajustadas dos índices de preços calculados pelo IBGE (Buzanovsky, Pinto e Cruz, 1995). Há muito que fazer nessa área e é encorajador que pesquisadores do IBGE tenham recentemente submetido ao CNPq projeto de pesquisa sobre o tema para complementar o financiamento do trabalho (Feijó, Silva e Carvalho, 1996).

Finalmente, vale ainda comentar que o valor dos dados produzidos pelas pesquisas pode ser bastante aumentado quando essas são combinadas entre si e com dados de outras fontes, mediante estruturas analíticas integradoras, tais como a das contas nacionais ou os relatórios integrados de indicadores sociais. Esses esquemas de análise permitem criticar e validar dados de forma que as pesquisas isoladas não contemplam, mas são ainda pouco explorados, diante de seu potencial.

5. Conclusões e Desafios

Esse artigo procurou rever a evolução observada no emprego de métodos para seleção de amostras, estimação e análise de dados amostrais complexos no IBGE, fazendo um paralelo com o desenvolvimento da teoria de amostragem em geral. O IBGE emerge dessa revisão como um usuário maduro de métodos para seleção e estimação de amostras, embora o mesmo não se possa dizer em termos da análise de resultados.

Há grandes desafios a superar para manter a tradição nas áreas em que o IBGE tem experiência adquirida, bem como para ampliar a utilização das modernas técnicas estatísticas disponíveis em benefício da redução de custos via amostras menores, aumento da eficiência e da precisão mediante revisão dos desenhos amostrais e estimadores empregados, e aumento da velocidade de obtenção dos resultados via emprego de sistemas genéricos de estimação tais como SUDAAN e GES. Há que ampliar o uso de técnicas estatísticas de análise de dados tais como modelos hierárquicos, modelos de séries temporais, etc. implantando uma cultura de análise dos resultados que realimente o planejamento de pesquisas futuras e contribua de maneira mais efetiva no debate informado dos grandes temas de interesse nacional.

Há também desafios em termos de ampliar ainda mais o acesso aos microdados e outros resultados das pesquisas, o principal deles sendo a divulgação mais rápida das informações. Os recentes progressos no emprego de tecnologias modernas como CD-ROM e redes como a

INTERNET para a disseminação das informações deverão provocar aumento da demanda e do número de usuários em contato com o IBGE, que precisará estar preparado para atender essa demanda de forma mais ágil e satisfatória.

Será cada vez mais importante que o IBGE divulgue junto ao público usuário as metodologias adotadas na produção dos dados, juntamente com informações sobre a qualidade dos mesmos, incluindo estimativas da precisão amostral, taxas de não-resposta e de imputação de itens, quando for o caso. Esse processo contribuirá para "educar" os usuários no entendimento mais pleno do processo de produção de informações empregando pesquisas por amostragem.

Vale mencionar ainda a necessidade de completar a "conquista" de grupos ainda não completamente adeptos do uso de amostragem, tanto dentro do IBGE como entre seus usuários menos crédulos.

Para vencer todos esses desafios é imprescindível que o IBGE recrute, forme e mantenha em seus quadros pessoal técnico qualificado, em particular nas áreas de Estatística e de amostragem. Sem competência no manejo do "telescópio da Estatística", ficará prejudicada a visão IBGEana da realidade brasileira.

6. Referências

Albieri, S., Bianchini, Z.M. e Cardoso, R.L. (1995). Pesquisa domiciliar sobre padrões de vida - planejamento da amostra. Rio de Janeiro: IBGE, Diretoria de Pesquisas.

Albieri, S. e Dias, A.J.R. (1994). Metodologia de expansão da amostra do censo demográfico de 1991: uma descrição resumida. Rio de Janeiro: IBGE, Divisão de Metodologia.

Andersson, C. e Nordberg, L. (1994). A method for variance estimation of non-linear functions of totals in surveys - theory and software implementation. Journal of Official Statistics, 10, 395-406.

Bianchini, Z.M. (1990). Projeções de população e o ajuste das estimativas das pesquisas domiciliares: solução ou problema? Indicadores IBGE, 9, 91-99.

Bianchini, Z.M. e Vieira, M. (1996). O planejamento da amostra da pesquisa de orçamentos familiares 95/96. Rio de Janeiro: IBGE, Departamento de Metodologia.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279-292.

Bryk, A.S. e Raudenbush, S.W. (1992). Hierarchical linear models. Newbury Park: Sage Publications, Advanced Quantitative Techniques in the Social Sciences Series, 1.

Buzanovsky, A.M., Pinto, L.M.C. e Cruz, M.M. (1995). Ajustamento sazonal nos índices de preços ao consumidor amplo. Rio de Janeiro: IBGE, Departamento de Índices de Preços.

Cabral, M.D.B. (1985). Plano de amostragem da PIA 1984. Rio de Janeiro: IBGE.

Cochran, W.G. (1977). Sampling Techniques, third edition. New York: John Wiley & Sons.

Deville, J.C. e Särndal, C.E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.

Estevao, V., Hidiroglou, M.A. e Särndal, C.E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. Journal of Official Statistics, 11, 181-204.

Farias, A.M.L. e Barbosa, M.T.S. (1995). Estratificação da população na pesquisa mensal do comércio. Artigo submetido para publicação na Revista Brasileira de Estatística.

Feijó, C.A., Silva, D.B.N., e Carvalho, P.G.M. (1996). Análise longitudinal de dados industriais: projeto integrado de pesquisa CNPq. Rio de Janeiro: IBGE, ENCE.

Fuller, W.A. (1986). PC-CARP. Ames: Statistical Laboratory, Iowa State University.

Garcia, R.M. (1986). Estimação em pequeno domínio. Rio de Janeiro: Instituto de Matemática Pura e Aplicada, Informes de Matemática série D, 020/86.

Goldstein, H. e Silver, R. (1989). Multilevel and multivariate models in survey analysis. Em Skinner, C.J., Holt, D. e Smith, T.M.F., eds, Analysis of complex surveys, Chichester, John Wiley & Sons, cap. 11.

Göttgens, R., Zellen, B., Odekerken, M. e Hofman, L. (1991). BASCULA version 1.0 - a weighting package under MS-DOS - user manual. Voorburg: Statistics Netherlands.

Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. The American Statistician, v. 40, 27-31.

Holt, D., Smith, T.M.F. e Winter, P.D. (1980). Regression analysis of data from complex surveys. Journal of the Royal Statistical Society, A, 143, 474-487.

IBGE (1981a). Metodologia da pesquisa nacional por amostra de domicílios na década de 70. Rio de Janeiro: Série Relatórios Metodológicos, v. 1.

IBGE (1981b). Plano de expansão da amostra de dados gerais da PIA 1981. Rio de Janeiro: Departamento de Coordenação de Métodos.

IBGE (1982a). Pesquisa especial da indústria, notas metodológicas. Rio de Janeiro: Departamento de Coordenação de Métodos.

IBGE (1982b). Plano amostral da PIA-81. Rio de Janeiro: Divisão de Amostragem S.2.

IBGE (1982c). Crítica quantitativa, qualitativa e fórmulas de cálculo do IPP; projeto Índice de Preços ao Produtor (documento preliminar). Rio de Janeiro: Departamento de Coordenação de Métodos.

IBGE (1983a). Metodologia da pesquisa mensal de emprego 1980. Rio de Janeiro: Série Relatórios Metodológicos, v. 2.

IBGE (1983b). Metodologia do censo demográfico de 1980. Rio de Janeiro: Série Relatórios Metodológicos, v. 4.

IBGE (1983c). Pesquisa industrial de 1981: tratamento das situações especiais durante a fase de expansão. Rio de Janeiro: Departamento de Coordenação de Métodos.

IBGE (1984). Plano de amostragem da PIA-83. Rio de Janeiro.

IBGE (1991a). Indicadores conjunturais da indústria e produção, emprego e salário. Rio de Janeiro: Diretoria de Pesquisas, Série Relatórios Metodológicos, v. 11.

IBGE (1991b). Pesquisa anual de comércio PAC. Rio de Janeiro: Série Relatórios Metodológicos, v. 12.

IBGE (1992). Pesquisa de orçamentos familiares, volume 3, aspectos de amostragem. Rio de Janeiro: Série Relatórios Metodológicos, v. 10.

IBGE (1994a). Sistema nacional de indicadores de preços ao consumidor: métodos de cálculo. Rio de Janeiro: Diretoria de Pesquisas, Série Relatórios Metodológicos, v. 14.

IBGE (1994b). Pesquisa de locais de compra PLC/1988. Rio de Janeiro: Departamento de Índices de Preços.

IBGE (1995). Aspectos de amostragem relativos à pesquisa cadastro de 1995 (versão preliminar). Rio de Janeiro: Departamento de Metodologia.

Menezes, A.C.F., Almeida, R.A.P. e Bianchini, Z.M. (1991). Revisão da PNAD: a questão amostral - módulo II do anteprojeto. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Série Textos para Discussão n. 54.

Menezes, A.C.F. et al. (1986). Pesquisa nacional por amostra de domicílios: redução do número de domicílios na amostra e proposta de novos estimadores. Rio de Janeiro: IBGE.

Moura, F.A.S. (1994). Small area estimation using multilevel models. Southampton: University of Southampton, Department of Social Statistics, tese de doutorado não publicada.

Muller, C.C., Silva, G. e Villalobos, A.G. (1988). Pesquisa agropecuária do Paraná - safra 1986-87 (programa de aperfeiçoamento das estatísticas agropecuárias). Revista Brasileira de Estatística, v. 49, 55-84.

Nathan, G. e Holt, D. (1980). The effect of survey design on regression analysis. Journal of the Royal Statistical Society, B, 42, 377-386.

Nathan, G. e Smith, T.M.F. (1989). The effect of selection on regression analysis. Em Skinner, C.J., Holt, D. e Smith, T.M.F., eds, Analysis of complex surveys, Chichester, John Wiley & Sons, cap. 7.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. International Statistical Review, 61, 317-337.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. Biometrika, 57, 377-387.

Särndal, C.E., Swensson, B. e Wretman, J. (1992). Model assisted survey sampling. New York: Springer-Verlag.

Sautory, O. (1993). La macro CALMAR: redressement d'un échantillon par calage sur marges. Paris: INSEE.

Shah, B.V., Barnwell, B.G., Hunt, P.N. e LaVange, L.M. (1992). SUDDAN user's manual - professional software for survey data analysis for multi-stage sample designs - release 6.0. North Carolina: Research Triangle Institute.

Silva, D.B.N. (1992). Aplicação de modelos lineares dinâmicos bayesianos em pesquisas repetidas no tempo. Rio de Janeiro: Universidade Federal do Rio de Janeiro, Instituto de Matemática e Estatística, tese de mestrado.

Silva, D.B.N. (1995). Modelling compositional time series from repeated surveys. Southampton: University of Southampton, Faculty of Mathematical Studies, relatório não publicado.

Silva, L.F. e Bianchini, Z.M. (1990). A redução da amostra e a utilização de duas frações amostrais no censo demográfico de 1990. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Série Textos para Discussão, n. 33.

Silva, P.L.N. (1989). Macros para seleção de amostras. Rio de Janeiro: IBGE, Núcleo de Metodologia.

Silva, P.L.N. (1996). Utilizing auxiliary information in sample survey estimation and analysis. Southampton: Department of Social Statistics, University of Southampton, tese de doutorado não publicada.

Silva, P.L.N. e Bianchini, Z.M. (1995). Data editing issues and strategies at the brazilian central statistical office. Artigo submetido para publicação no Journal of Official Statistics.

Silva, P.L.N., Bianchini, Z.M. e Albieri, S. (1993). Uma proposta de metodologia para a expansão da amostra do censo demográfico de 1991. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Série Textos para Discussão, n. 62.

Silva, P.L.N. e Moura, F.A.S. (1988). Redução da amostra da pesquisa mensal de emprego: estratégia para reduzir o custo da pesquisa. Revista Brasileira de Estatística, v. 49, 65-95.

Silva, P.L.N. e Moura, F.A.S. (1990). Efeito de conglomeração da malha setorial do censo demográfico de 1980. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Série Textos para Discussão n. 32.

Silva, P.L.N., Silva, D.B.N., Moura, F.A.S. e Jooris, L.R. (1995). Frame problems and survey design for the Brazilian annual retail and wholesale trade survey. Artigo submetido para publicação na revista Estadística.

Silva, P.L.N. e Skinner, C.J. (1995). Estimating distribution functions with auxiliary information using poststratification. Journal of Official Statistics, 11, 277-294.

Silva, P.L.N. e Skinner, C.J. (1996). Variable selection for regression estimation in finite populations. Artigo submetido para publicação no Journal of the American Statistical Association.

Skinner, C.J., Holt, D. e Smith, T.M.F. (1989). Analysis of complex surveys. Chichester: John Wiley & Sons.

Smith, T.M.F. (1988). To weight or not to weight, that is the question. Em Bernardo, J.M., DeGroot, M.H., Lindley, D.V. e Smith, A.F.M., eds, Bayesian Statistics 3, Oxford, Oxford University Press, 437-451.

Smith, T.M.F. (1989). Introduction to part B. Em Skinner, C.J., Holt, D. e Smith, T.M.F., eds, Analysis of complex surveys, Chichester, John Wiley & Sons, p. 146-147.

Smith, T.M.F. (1993). Populations and selection: limitations of statistics. Journal of the Royal Statistical Society, A, 156, 145-164.

Vasconcellos, M.T.L. (1983) Metodologia do estudo nacional da despesa familiar - ENDEF: objetivos, descrição e metodologia usadas no ENDEF. Rio de Janeiro: IBGE, Departamento de Coordenação de Métodos.

Westat (1995). A user's guide to WesVarPC, version 1.0. Rockville: Westat.

Wolter, K.M. (1985). Introduction to variance estimation. New York: Springer-Verlag.